

## RERANDOMIZATION TO IMPROVE COVARIATE BALANCE IN EXPERIMENTS<sup>1</sup>

BY KARI LOCK MORGAN AND DONALD B. RUBIN

*Duke University and Harvard University*

Randomized experiments are the “gold standard” for estimating causal effects, yet often in practice, chance imbalances exist in covariate distributions between treatment groups. If covariate data are available before units are exposed to treatments, these chance imbalances can be mitigated by first checking covariate balance *before* the physical experiment takes place. Provided a precise definition of imbalance has been specified in advance, unbalanced randomizations can be discarded, followed by a rerandomization, and this process can continue until a randomization yielding balance according to the definition is achieved. By improving covariate balance, rerandomization provides more precise and trustworthy estimates of treatment effects.

**1. A brief history of rerandomization.** Randomized experiments are the “gold standard” for estimating causal effects, because randomization balances all potential confounding factors *on average*. However, if in a particular experiment, a randomization creates groups that are notably unbalanced on important covariates, should we proceed with the experiment, rather than rerandomizing and conducting the experiment on balanced groups?

With  $k$  independent covariates, the chance of *at least one* covariate showing a “significant difference” between treatment and control groups, at significance level  $\alpha$ , is  $1 - (1 - \alpha)^k$ . For a modest 10 covariates and a 5% significance level, this probability is 40%. “Most experimenters on carrying out a random assignment of plots will be shocked to find how far from equally the plots distribute themselves” [Fisher (1926)]. The danger of relying on pure randomization to balance covariates has been described in Seidenfeld (1981); Urbach (1985); Krause and Howard (2003); Rosenberger and Sverdlov (2008); Rubin (2008a); Keele et al. (2009) and Worrall (2010). Also, there exists much discussion historically over whether randomization should be preferred over a purposefully balanced assignment [Gosset (1938); Yates (1939); Greenberg (1951); Harville (1975); Arnold (1986); Kempthorne (1986)]. Our view is that with rerandomization, we can retain the advantages of randomization, while also ensuring balance.

---

Received November 2011; revised April 2012.

<sup>1</sup>Supported in part by the following Grants: NSF SES-0550887, NSF IIS-1017967 and NIH R01DA23879.

*MSC2010 subject classifications.* 62K99.

*Key words and phrases.* Randomization, treatment allocation, experimental design, clinical trial, causal effect, Mahalanobis distance, Hotelling’s  $T^2$ .

It is standard in randomized experiments today to collect covariate data and check for covariate balance, yet typically this is done after the experiment has started. If covariate data are available before the physical experiment has started, a randomization should be checked for balance *before* the physical experiment is conducted. If lack of balance is noted, as Gosset stated, “it would be pedantic to continue with an arrangement of plots known beforehand to be likely to lead to a misleading conclusion” [Gosset (1938)]. It appears that Fisher would agree. In Rubin (2008a), Rubin recounts the following conversation with his advisor Bill Cochran:

*Rubin:* What if, in a randomized experiment, the chosen randomized allocation exhibited substantial imbalance on a prognostically important baseline covariate?

*Cochran:* Why didn’t you block on that variable?

*Rubin:* Well, there were many baseline covariates, and the correct blocking wasn’t obvious; and I was lazy at that time.

*Cochran:* This is a question that I once asked Fisher, and his reply was unequivocal:

*Fisher (recreated via Cochran):* Of course, if the experiment had not been started, I would rerandomize.

A similar conversation between Fisher and Savage, wherein Fisher advocates rerandomization when faced with an undesirable randomization, is documented in Savage [(1962), page 88].

Checking covariates and rerandomizing when needed for balance has been advocated repeatedly. Sprott and Farewell (1993) recommend rerandomization when “obvious” lack of balance is observed. Rubin (2008a) suggests that if “important imbalances exist, rerandomize, and continue to do so until satisfied.” For clinical trials, Worrall (2010) states that “if such baseline imbalances are found then the recommendation . . . is to re-randomize in the hope that this time no baseline imbalances will occur.” Cox (2009) and Bruhn and McKenzie (2009) have advocated rerandomization, suggesting either to do multiple randomizations and pick the “best,” or to specify a bound for the difference in treatment and control covariate means for each covariate, following the “Big Stick” method of Soares and Wu (1985), and rerandomize until all differences are within these bounds. The latter rerandomization method was used in Maclure et al. (2006).

There are also many sources giving reasons *not* to rerandomize. Good accounts of the debate over rerandomization can be found in Urbach (1985) and Raynor (1986). The most common critique of rerandomization is that forms of analysis utilizing Gaussian distribution theory are no longer valid [Fisher (1926); Anscombe (1948a); Grundy and Healy (1950); Holschuh (1980); Bailey (1983); Urbach (1985); Bailey (1986); Bailey and Rowley (1987)]. Rerandomization changes the distribution of the test statistic, most notably by decreasing the true standard error, thus traditional methods of analysis that do not take this into account will result in overly “conservative” inferences in the sense that tests will reject true null hypotheses less often than the nominal level and confidence intervals will cover the

true value more often than the nominal level. However, randomization-based inference is still valid [Anscombe (1948a); Kempthorne (1955); Brillinger, Jones and Tukey (1978); Tukey (1993); Rosenberger and Lachin (2002); Moulton (2004)], because the rerandomization can be accounted for during analysis.

All other critiques of rerandomization, of which we are aware, deal with “ad-hoc” rerandomization, that is, rejecting randomizations without specifying a rejection criterion in advance. We only advocate rerandomization if the decision to rerandomize or not is based on a pre-specified criterion. By specifying an objective rerandomization rule before randomizing, and then analyzing results using randomization-based methods, we can, in most circumstances, finesse all existing criticisms of rerandomizing.

Some may think that rerandomization is unnecessary with large sample sizes, because as the sample size increases, the difference in covariate means between groups gets smaller, essentially proportional to the square root of the sample size. However, at the same rate, confidence intervals and significance tests are getting more sensitive to small differences in outcome means, which can be driven by small differences in covariate means.

Despite the ongoing discussion about rerandomization and the fact that it is widely used in practice [Holschuh (1980); Urbach (1985); Bailey and Rowley (1987); Imai, King and Stuart (2008); Bruhn and McKenzie (2009)], little has been published on the mathematical implications of rerandomization. Remarkably, it appears that no source even makes explicit the conditions under which rerandomization is valid. Although a few rerandomization methods have been proposed [Moulton (2004); Maclure et al. (2006); Bruhn and McKenzie (2009); Cox (2009)], the implications have not been theoretically explored, to the best of our knowledge. The only published theoretical results accompanying a rerandomization procedure appear to be those in Cox (1982), which proposed rerandomization to lower the sampling variance of covariance-adjusted estimates. Here we aim to fill these lacuna by (a) making explicit the sufficient conditions under which rerandomization is valid, (b) describing in detail a principled procedure for implementing rerandomization and (c) providing corresponding theoretical results.

## 2. Rerandomization in general.

2.1. *Procedure.* The procedure for implementing rerandomization is depicted in Figure 1, and has the following steps:

- (1) Collect covariate data.
- (2) Specify a balance criterion determining when a randomization is acceptable.
- (3) Randomize the units to treatment groups.
- (4) Check the balance criterion; if the criterion is met, go to Step (5). Otherwise, return to Step (3).

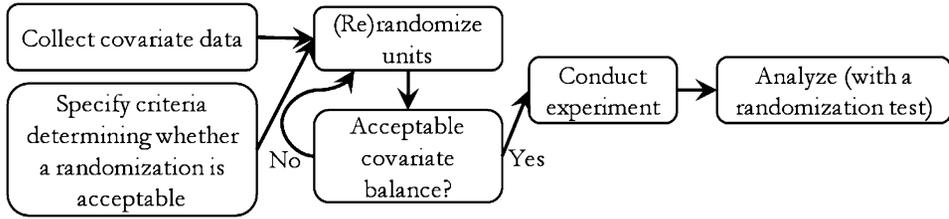


FIG. 1. Procedure for implementing rerandomization.

- (5) Conduct the experiment using the final randomization obtained in Step (4).
- (6) Analyze the results using a randomization test, keeping only simulated randomizations that satisfy the balance criterion specified in Step (2).

Let  $\mathbf{x}$  be the  $n \times k$  covariate matrix representing  $k$  covariates measured on  $n$  experimental units. Here we assume that a sample of units has already been selected and is fixed. Because we are not considering the sampling mechanism, we are only interested in the extent to which a causal effect estimate obtained in *this* randomized experiment is a good estimate of the true causal effect within the selected sample. The  $\mathbf{x}$  matrix includes all the observed covariates for which balance between groups is desired, which may include original covariates, and any functions of original covariates, such as transformations, powers and interactions. Let  $\mathbf{W}$  be the  $n$ -dimensional treatment assignment vector indicating the treatment group for each unit. The rerandomization criterion is based on a row-exchangeable scalar function of  $\mathbf{x}$  and  $\mathbf{W}$ .

$$\varphi(\mathbf{x}, \mathbf{W}) = \begin{cases} 1, & \text{if } \mathbf{W} \text{ is an acceptable randomization,} \\ 0, & \text{if } \mathbf{W} \text{ is not an acceptable randomization.} \end{cases}$$

The function  $\varphi$  can vary depending on the relative importance of balancing different covariates, on the level of covariate balance desired and on the computational power available, but it is specified in advance.

More generally, we can define a set of acceptance criteria,  $\mathcal{S} = \{\varphi_s\}$ , from which we choose at each step,  $s$ , either deterministically or stochastically, where this choice can depend on the step, so that, for example, we can become more lenient as the steps increase without success. In this more general situation,  $\varphi_s(\mathbf{x}, \mathbf{W})$  denotes the acceptance criterion for step  $s$ . Although our theoretical results in Sections 2 and 3 hold for this more general setup, in practice we expect that the common choice will be one function for all steps, and to avoid notational clutter, we present results with one criterion.

Once  $\varphi$  has been specified, units are randomized to treatment groups (Step 3). In the simplest form of rerandomization, this can be done with no restrictions; for example, randomly choose an assignment vector  $\mathbf{W}$  from all possible vectors, or equivalently from all possible partitions of the units into groups. In practice, the initial randomization is typically done with some restriction to equalize treatment group sizes.

Rerandomization is simply a tool that allows us to draw from some predetermined set of acceptable randomizations,  $\{\mathbf{W} \mid \varphi(\mathbf{x}, \mathbf{W}) = 1\}$ . Rerandomization is analogous to rejection sampling; a way to draw from a set that may be tedious to enumerate. Specifying a set of acceptable randomizations and then choosing randomly from this set is recommended by [Kempthorne \(1955, 1986\)](#) and [Tukey \(1993\)](#), and [Moulton \(2004\)](#) notes that rerandomization may be required for implementation of this idea when the set of acceptable randomizations is difficult to enumerate a priori.

Within this framework, rerandomization simply generalizes classical experimental designs. For the basic completely randomized experiment with fixed sample sizes in each treatment group,  $\varphi(\mathbf{x}, \mathbf{W}) = 1$  when the number of units assigned to each group matches the predetermined group sizes. For a randomized block experiment,  $\varphi(\mathbf{x}, \mathbf{W}) = 1$  when predetermined numbers of units within each block are assigned to each treatment group. For a Latin square,  $\varphi(\mathbf{x}, \mathbf{W}) = 1$  when the randomization satisfies the Latin square design. These classical designs can be readily sampled from, so rerandomization is computationally inefficient, although equivalent, but for other functions,  $\varphi$ , rerandomization may be a more straightforward technique. Rerandomization can also be used together with any classical design. For example, in a medical experiment on hypertensive drugs, we may block on sex and a coarse categorization of baseline blood pressure, and use rerandomization to balance the remaining covariates, including fine baseline blood pressure.

Researchers are free to choose any function  $\varphi$ , provided it is chosen in advance. [Section 2.3](#) describes the conditions necessary to maintain general unbiasedness of simple point estimation, [Section 3](#) recommends a particular class of functions and studies theoretical properties of this choice and [Section 4](#) discusses some reasons for choosing an affinely invariant  $\varphi$ .

*2.2. Analysis by randomization tests.* Under most forms of rerandomization, increasing balance in the covariates will typically create more precise estimated treatment effects, making traditional Gaussian distribution-based forms of analysis statistically too conservative. However, the final data can be analyzed using a randomization test, maintaining valid frequentist properties. As Fisher stated, “It seems to have escaped recognition that the physical act of randomization . . . affords the means, in respect of any particular body of data, of examining a wider hypothesis in which no normality of distribution is implied” [[Fisher \(1935\)](#)]. This physical act of randomization need not be pure randomization, but any randomization scheme that can be replicated when conducting the randomization test.

We are interested in the effect of treatment assignment,  $\mathbf{W}$ , on an outcome,  $\mathbf{y}$ . Let  $y_i(W_i)$  denote the  $i$ th unit's,  $\{i = 1, \dots, n\}$ , potential outcome under treatment assignment  $W_i$ , following the Rubin causal model [[Rubin \(1974\)](#)]. Although rerandomization can be applied to any number of treatment conditions, to convey essential ideas most directly, we consider only two, and refer to these conditions

as treatment and control. Let

$$W_i = \begin{cases} 1, & \text{if treated,} \\ 0, & \text{if control.} \end{cases}$$

Let  $\mathbf{Y}_{obs}(\mathbf{W})$  denote the vector of observed outcome values:

$$(1) \quad Y_{obs,i} = y_i(1)W_i + y_i(0)(1 - W_i),$$

where for notational simplicity the subscript *obs* means  $obs(\mathbf{W})$ . Under the sharp null hypothesis of no treatment effect on any unit,  $y_i(1) = y_i(0)$  for every  $i$ , and thus the vector  $\mathbf{Y}_{obs}$  is the same for every treatment assignment  $\mathbf{W}$ . Consequently, leaving  $\mathbf{Y}_{obs}$  fixed and simulating many acceptable randomization assignments,  $\mathbf{W}$ , we can empirically create the distribution of any estimator,  $g(\mathbf{x}, \mathbf{W}, \mathbf{y}_{obs})$ , if the null hypothesis were true. To account for the rerandomization, each simulated randomization must also satisfy  $\varphi(\mathbf{x}, \mathbf{W}) = 1$ . Once the desired number of randomizations has been simulated, the proportion of simulated randomizations with estimated treatment effect as extreme or more extreme than that observed in the experiment is the  $p$ -value. Although a full permutation test (including all the acceptable randomizations) is necessary for an exact  $p$ -value, the number of simulated randomizations can be increased to provide a  $p$ -value with any desired level of accuracy. This test can incorporate whatever rerandomization procedure was used, will preserve the significance level of the test [Moulton (2004)] and works for any estimator. Brillinger, Jones and Tukey (1978), Tukey (1993) and Rosenberger and Lachin [(2002), Chapter 7] suggest using randomization tests to assess significance when restricted randomization schemes are used.

Because analysis by a randomization test requires generating many acceptable randomizations, computational time can be important to consider in advance. Define  $p_a \equiv P(\varphi = 1)$  to be the proportion of acceptable randomizations. The choice of  $p_a$  involves a trade-off between better balance and computational time; smaller values of  $p_a$  ensure better balance, but they also imply a longer expected waiting time to obtain an acceptable randomization, at least without clever computational devices. The number of randomizations required to get one acceptable randomization follows a geometric distribution with parameter  $p_a$ , so  $N$  simulated acceptable randomizations for a randomization test will require on average  $N/p_a$  randomizations to be generated.

The chosen  $p_a$  must leave enough acceptable randomizations to perform a randomization test. In practice this is rarely an issue, because the number of possible randomizations is huge even for modest  $n$ . To illustrate, the number of possible randomizations for  $n = \{30, 50, 100\}$  randomizing to two equally sized treatment groups,  $\binom{n}{n/2}$ , is on the order of  $\{10^8, 10^{14}, 10^{29}\}$ , respectively. However, for small sample sizes, care should be taken to ensure the number of acceptable randomizations does not become too small, for example, less than 1000.

By employing the duality between confidence intervals and tests, for additive treatment effects a confidence interval can be produced from a randomization distribution as the set of all values for which the observed data would not reject

such a null hypothesized value [Lehmann and Romano (2005); Manly (2007), Section 3.5, Section 1.4]. Garthwaite (1996) provides an efficient algorithm for generating a confidence interval for additive effects from a randomization test. The assumption of additivity is statistically conservative, at least asymptotically, as implied by Neyman’s [Splawa-Neyman (1990)] results on standard errors being overestimated when assuming it relative to the actual standard errors. A randomization test can be applied to any sharp null hypothesis, that is, a hypothesis such that the observed data implies specific values for all missing potential outcomes.

When the covariates being balanced are correlated with the outcome variable, then rerandomization increases precision (Section 3.2). A randomization test reflects this increase in precision. Standard asymptotic-based frequentist analysis procedures that do not take the rerandomization into account will be statistically conservative. Not only will distribution-based standard errors not incorporate the increase in precision, but the act of rerandomizing itself will increase the *estimated* standard error beyond that of pure randomization. If the total variance in the outcome is fixed, decreasing the actual sampling variance between treatment group means (by ensuring better balance), increases the variance within groups, and it is this variance within groups that is traditionally used to estimate the standard error [Fisher (1926)]. Thus, although rerandomization decreases the *true* standard error, it actually increases the standard error as estimated by traditional methods. For both of these reasons, the regular estimated standard errors will overestimate the true standard error, and using the corresponding distribution-based methods of analysis after rerandomization results in overly wide confidence intervals and less powerful tests of hypotheses.

2.3. *Maintaining an unbiased estimate.* Although not needed to motivate rerandomization, we assume one goal is to estimate the average treatment effect

$$(2) \quad \begin{aligned} \tau &\equiv \overline{y(1)} - \overline{y(0)} \\ &= \frac{\sum_{i=1}^n y_i(1)}{n} - \frac{\sum_{i=1}^n y_i(0)}{n}. \end{aligned}$$

The fundamental problem in causal inference is that, because we only observe  $y_i(W_i)$  for each unit, we cannot calculate (2) directly, and we must estimate  $\tau$  using only  $\mathbf{Y}_{obs}$ . In this section, we assume the Stable Unit Treatment Value Assumption (SUTVA) [Rubin (1980)]: the potential outcomes are fixed and do not change with different possible assignment vectors  $\mathbf{W}$ .

The average treatment effect  $\tau$  is usually estimated by the difference in observed sample means,

$$\hat{\tau} \equiv \overline{Y}_{obs,T} - \overline{Y}_{obs,C},$$

where

$$\overline{Y}_{obs,T} \equiv \frac{\sum_{i=1}^n W_i y_i(1)}{\sum_{i=1}^n W_i} \quad \text{and} \quad \overline{Y}_{obs,C} \equiv \frac{\sum_{i=1}^n (1 - W_i) y_i(0)}{\sum_{i=1}^n (1 - W_i)}.$$

**THEOREM 2.1.** *Suppose  $\sum_{i=1}^n W_i = \sum_{i=1}^n (1 - W_i)$  and  $\varphi(\mathbf{x}, \mathbf{W}) = \varphi(\mathbf{x}, \mathbf{1} - \mathbf{W})$ ; then  $\mathbb{E}(\hat{\tau} \mid \mathbf{x}, \varphi = 1) = \tau$ .*

**PROOF.** Under the specified conditions,  $\mathbf{W}$  and  $\mathbf{1} - \mathbf{W}$  are exchangeable. Therefore, after rerandomization  $\mathbb{E}(W_i \mid \mathbf{x}, \varphi = 1) = \mathbb{E}(1 - W_i \mid \mathbf{x}, \varphi = 1) \forall i$ , so  $\mathbb{E}(W_i \mid \mathbf{x}, \varphi = 1) = \mathbb{E}(1 - W_i \mid \mathbf{x}, \varphi = 1) = 1/2 \forall i$ . Hence

$$\begin{aligned} \mathbb{E}(\hat{\tau} \mid \mathbf{x}, \varphi = 1) &= \mathbb{E}\left(\frac{\sum_{i=1}^n W_i Y_{i,obs}}{n/2} - \frac{\sum_{i=1}^n (1 - W_i) Y_{i,obs}}{n/2} \mid \mathbf{x}, \varphi = 1\right) \\ &= \mathbb{E}\left(\frac{\sum_{i=1}^n W_i y_i(1)}{n/2} - \frac{\sum_{i=1}^n (1 - W_i) y_i(0)}{n/2} \mid \mathbf{x}, \varphi = 1\right) \\ &= \frac{\sum_{i=1}^n \mathbb{E}(W_i \mid \varphi = 1) y_i(1)}{n/2} - \frac{\sum_{i=1}^n (1 - \mathbb{E}(W_i \mid \mathbf{x}, \varphi = 1)) y_i(0)}{n/2} \\ &= \frac{\sum_{i=1}^n (1/2) y_i(1)}{n/2} - \frac{\sum_{i=1}^n (1/2) y_i(0)}{n/2} \\ &= \tau. \end{aligned} \quad \square$$

Theorem 2.1 holds for all outcome variables. Corollary 2.2 follows by the same logic.

**COROLLARY 2.2.** *If  $\sum_{i=1}^n W_i = \sum_{i=1}^n (1 - W_i)$  and  $\varphi(\mathbf{x}, \mathbf{W}) = \varphi(\mathbf{x}, \mathbf{1} - \mathbf{W})$ , then  $\mathbb{E}(\overline{V}_T - \overline{V}_C \mid \mathbf{x}, \varphi = 1) = 0$  for any observed or unobserved covariate  $V$ .*

If sample sizes are not fixed in advance, but each unit has  $E(W_i \mid \mathbf{x}) = 1/2$  in the initial randomization,  $\hat{\tau}$  is only necessarily an unbiased estimate under the assumption of additivity. As a small example under nonadditivity, consider  $\mathbf{x} = (0, 1, 2)$ ,  $\mathbf{y}(1) = (1, 1, 0)$  and  $\mathbf{y}(0) = (0, 0, 1)$ . When  $\varphi(\mathbf{x}, \mathbf{W}) = 1$  if the difference in  $x$  means between the two groups is 0 and  $\varphi = 0$  otherwise, the only two acceptable randomizations are  $\mathbf{W} = (0, 1, 0)$  and  $\mathbf{W} = (1, 0, 1)$ . For either acceptable randomization,  $\hat{\tau} = 1/2$ , yet  $\tau = 1/3$ . This artificial example also illustrates that if the treatment groups are of unequal size,  $\hat{\tau}$  will not necessarily be an unbiased estimate after rerandomization. If the treatment group includes two units and the control group one unit, and  $\varphi(\mathbf{x}, \mathbf{W})$  is the same as before, then the only acceptable randomization is  $\mathbf{W} = (1, 0, 1)$ , and once again,  $\hat{\tau} = 1/2$ , whereas  $\tau = 1/3$ .

**3. Rerandomization using Mahalanobis distance.** To simplify the statement of theoretical results, we assume the sample sizes for the treatment and control groups are fixed in advance, with  $p_w$  the fixed proportion of treated units,

$$(3) \quad p_w = \frac{\sum_{i=1}^n W_i}{n}.$$

Let  $\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C$  be the  $k$ -dimensional vector of the difference in covariate means between the treatment and control groups,

$$(4) \quad \bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C = \frac{\mathbf{x}'\mathbf{W}}{np_w} - \frac{\mathbf{x}'(\mathbf{1} - \mathbf{W})}{n(1 - p_w)} = \frac{\mathbf{x}'(\mathbf{W} - p_w\mathbf{1})}{np_w(1 - p_w)}.$$

We consider Mahalanobis distance as a rerandomization criterion because it is an affinely invariant scalar measure of multivariate covariate balance. Mahalanobis distance is defined by

$$(5) \quad M \equiv (\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)' [\text{cov}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)]^{-1} (\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)$$

$$(6) \quad = np_w(1 - p_w) (\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)' \text{cov}(\mathbf{x})^{-1} (\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C),$$

where  $\text{cov}(\mathbf{x})$  represents the sample covariance matrix of  $\mathbf{x}$ . The quantities  $n$ ,  $p_w$  and  $\text{cov}(\mathbf{x})$  are known and constant across randomizations. If  $\text{cov}(\mathbf{x})$  is singular, for example, if  $k \geq n$ , then  $\text{cov}(\mathbf{x})^{-1}$  can be replaced with the pseudo-inverse of  $\text{cov}(\mathbf{x})$ . For cluster randomized experiments, see Hansen and Bowers (2008).

Due to the finite population central limit theorem,  $\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C$  is asymptotically multivariate normally distributed over its randomization distribution [Erdős and Rényi (1959); Hájek (1960)]. Normality of  $\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C$  is not necessary for rerandomization, but assuming normality allows for the theoretical results of this section. If  $\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C$  is multivariate normal, then under pure randomization,  $M \sim \chi_k^2$  [Mardia, Kent and Bibby (1980), page 62];  $M$  is the statistic used in Hotelling’s  $T^2$  test, but note that here  $M$  follows a  $\chi_k^2$  distribution because  $\mathbf{x}$  is considered fixed.

A randomization is deemed “acceptable” whenever  $M$  falls below a certain threshold,  $a$ . Let  $p_a$  be the proportion of randomizations that are acceptable, so that  $P(M \leq a) = p_a$ . Either  $a$  or  $p_a$  can be specified a priori, and then the other is fixed either using  $M \sim \chi_k^2$  if sample sizes are large enough or using an empirical distribution of  $M$  achieved through simulation. The rerandomization criterion,  $\varphi_M$ , is

$$(7) \quad \varphi_M(\mathbf{x}, \mathbf{W}) \equiv \begin{cases} 1, & \text{if } M \leq a, \\ 0, & \text{if } M > a. \end{cases}$$

3.1. Covariate balance under  $\varphi_M$ .

**THEOREM 3.1.** Assume rerandomization is conducted using  $\varphi_M$  with  $p_w = 1/2$ , and the covariate means are multivariate normal; then

$$(8) \quad \text{cov}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C \mid \mathbf{x}, \varphi_M = 1) = v_a \text{cov}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C \mid \mathbf{x}, ),$$

where

$$(9) \quad v_a \equiv \frac{2}{k} \times \frac{\gamma(k/2 + 1, a/2)}{\gamma(k/2, a/2)} = \frac{P(\chi_{k+2}^2 \leq a)}{P(\chi_k^2 \leq a)}$$

and  $\gamma$  denotes the incomplete gamma function:  $\gamma(b, c) \equiv \int_0^c y^{b-1} e^{-y} dy$ .

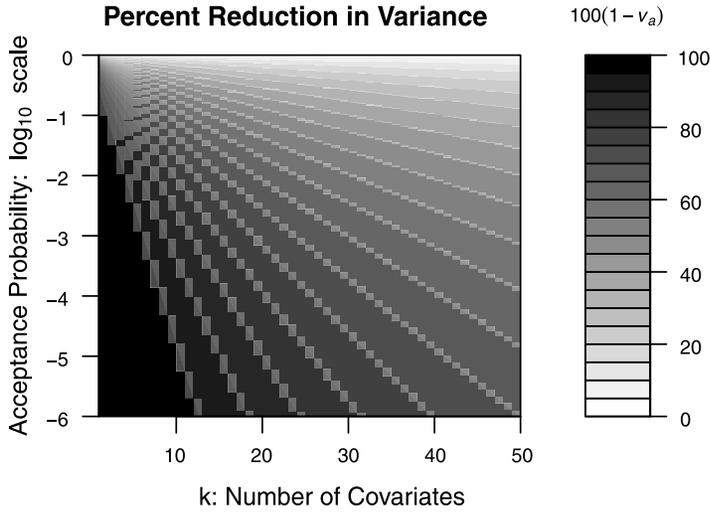


FIG. 2. The percent reduction in variance for each covariate difference in means, as a function of the number of covariates and the proportion of randomizations accepted.

The proof of Theorem 3.1 is in the [Appendix](#).

In the field of matching, emphasis has been placed on “percent reduction in bias” [Cochran and Rubin (1973)]. In the context of randomized experiments there is no bias, and rerandomization instead reduces the sampling variance of the difference in covariate means, yielding differences that are more closely concentrated around 0. Define the *percent reduction in variance*, the percentage by which rerandomization reduces the randomization variance of the difference in means, for each covariate,  $x_j$ , by

$$(10) \quad 100 \left( \frac{\text{var}(\bar{X}_{j,T} - \bar{X}_{j,C} | \mathbf{x}) - \text{var}(\bar{X}_{j,T} - \bar{X}_{j,C} | \mathbf{x}, \varphi = 1)}{\text{var}(\bar{X}_{j,T} - \bar{X}_{j,C} | \mathbf{x})} \right).$$

By Theorem 3.1, the percent reduction in variance for each covariate, and for any linear combination of these covariates, is

$$(11) \quad 100(1 - v_a)$$

and is shown as a function of  $k$  and  $p_a$  in Figure 2, where by (9),  $0 \leq v_a \leq 1$ . The lower the acceptance probability and the fewer covariates being balanced, the larger the percent reduction in variance.

Notice that Theorem 3.1 holds for any covariate distribution, as long as the sample size is large enough for the central limit theorem to ensure normally distributed covariate means. An exact value is not needed, and an estimate is used only to guide the choice of  $p_a$ ; it has no influence on the validity of resulting inferences.

3.2. *Precision of the estimated treatment effect.* Rerandomization improves precision, provided the outcome and covariates are correlated. Thus researchers can increase the power of tests and decrease the width of confidence intervals simply at the expense of computational time.

**THEOREM 3.2.** *If (a) rerandomization is conducted using  $\varphi_M$  with  $p_w = 1/2$ , (b) the covariate and outcome means are normally distributed, and (c) the treatment effect is additive, then the percent reduction in variance of  $\hat{\tau}$  is*

$$(12) \quad 100(1 - v_a)R^2,$$

where  $R^2$  represents the squared multiple correlation between  $\mathbf{y}$  and  $\mathbf{x}$  within a treatment group and  $v_a$  is as defined in (9).

**PROOF.** Regardless of the true relationship between the outcome and covariates, by additivity we can write

$$(13) \quad y_i(W_i) \mid \mathbf{x}_i = \beta_0 + \beta' \mathbf{x}_i + \tau W_i + e_i,$$

where  $\beta_0 + \beta' \mathbf{x}_i$  is the projection of  $y_i$  onto the space spanned by  $(\mathbf{1}, \mathbf{x})$ , and  $e_i$  is a residual that encompasses any deviations from the linear model. Then the estimated treatment effect,  $\hat{\tau}$ , can be expressed as

$$(14) \quad \hat{\tau} = \beta'(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C) + \tau + (\bar{e}_T - \bar{e}_C).$$

Because  $\tau$  is constant and the first and last terms are uncorrelated, we can express the variance of  $\hat{\tau}$  as

$$(15) \quad \begin{aligned} \text{var}(\hat{\tau}) &= \text{var}(\beta'(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)) + \text{var}(\bar{e}_T - \bar{e}_C) \\ &= \beta' \text{cov}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C)\beta + \text{var}(\bar{e}_T - \bar{e}_C). \end{aligned}$$

By Theorem 3.1, rerandomization modifies the first term by the factor  $v_a$ . Because under normality, orthogonality implies independence, the difference in residual means is independent of the difference in covariate means, and thus rerandomization has no affect on the second term. Therefore, the variance of  $\hat{\tau}$  after rerandomization restricting  $M \leq a$  is

$$(16) \quad \begin{aligned} \text{var}(\hat{\tau} \mid \mathbf{x}, M \leq a) &= \beta' \text{cov}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C \mid \mathbf{x}, M \leq a)\beta + \text{var}(\bar{e}_T - \bar{e}_C \mid \mathbf{x}, M \leq a) \\ &= v_a \beta' \text{cov}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C \mid \mathbf{x})\beta + \text{var}(\bar{e}_T - \bar{e}_C \mid \mathbf{x}). \end{aligned}$$

Let  $\sigma_e^2$  be the variance of the residuals and  $\sigma_y^2$  be the variance of the outcome within each treatment group, where  $\sigma_e^2 = \sigma_y^2(1 - R^2)$ . Thus

$$(17) \quad \text{var}(\bar{e}_T - \bar{e}_C \mid \mathbf{x}) = \frac{\sigma_e^2}{np_w(1 - p_w)} = \frac{\sigma_y^2(1 - R^2)}{np_w(1 - p_w)},$$

and

$$\begin{aligned}
 \beta' \text{cov}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C | \mathbf{x})\beta &= \text{var}(\hat{\tau} | \mathbf{x}) - \text{var}(\bar{e}_T - \bar{e}_C | \mathbf{x}) \\
 (18) \qquad \qquad \qquad &= \frac{\sigma_y^2 - \sigma_e^2}{np_w(1 - p_w)} \\
 &= \frac{\sigma_y^2 R^2}{np_w(1 - p_w)}.
 \end{aligned}$$

Therefore by (16), (17) and (18), the variance of  $\hat{\tau}$  after rerandomization is

$$\begin{aligned}
 \text{var}(\hat{\tau} | \mathbf{x}, M \leq a) &= v_a \beta' \text{cov}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C | \mathbf{x})\beta + \text{var}(\bar{e}_T - \bar{e}_C | \mathbf{x}) \\
 &= \frac{v_a \sigma_y^2 R^2}{np_w(1 - p_w)} + \frac{\sigma_y^2(1 - R^2)}{np_w(1 - p_w)} \\
 &= (1 - (1 - v_a)R^2) \frac{\sigma_y^2}{np_w(1 - p_w)} \\
 &= (1 - (1 - v_a)R^2) \text{var}(\hat{\tau} | \mathbf{x}).
 \end{aligned}$$

Thus the percent reduction in variance is  $100(1 - (1 - (1 - v_a)R^2)) = 100(1 - v_a)R^2$ .  $\square$

The percent reduction in variance for the estimated treatment effect, shown as a function of  $k$ ,  $p_a$  and  $R^2$  in Figure 3, is simply the percent reduction in variance for each covariate, scaled by  $R^2$ . Because under the specified conditions  $\hat{\tau}$  is unbiased

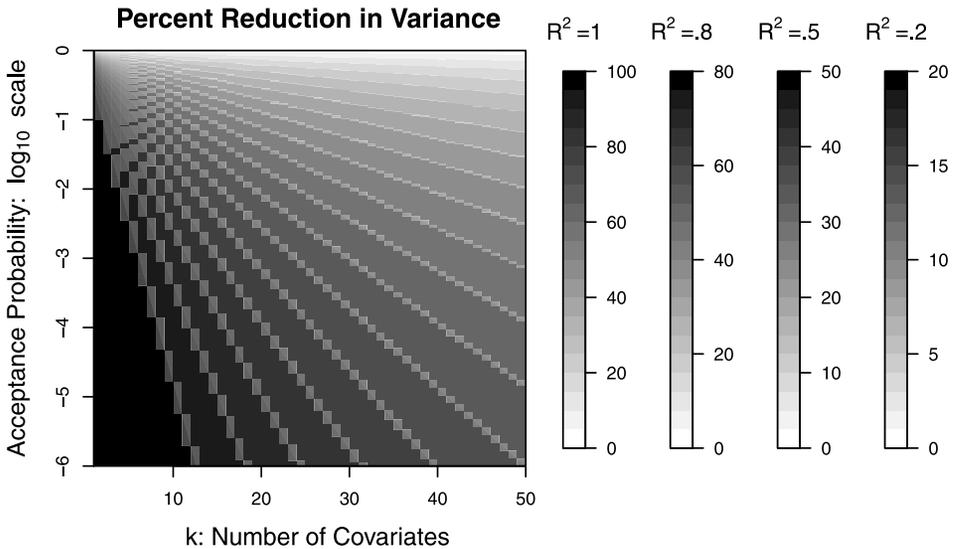


FIG. 3. The percent reduction in variance for the estimated treatment effect, as a function of the acceptance probability, the number of covariates, and  $R^2$ .

by Theorem 2.1,  $100(1 - v_a)R^2$  is not only the percent reduction in variance in the estimated treatment effect, but also the percent reduction in mean square error (MSE).

If regression (i.e., analysis of covariance) is used to adjust for imbalance in a completely randomized experiment, the percent reduction in variance is

$$(19) \quad 100 \left[ \left( 1 + \frac{M}{n} \right) R^2 - \frac{M}{n} \right]$$

[Cox (1982)], where  $M$  is as in (6). Comparing (19) to (10), we see that rerandomization can increase precision more than regression adjustment because there is no estimation of regression coefficients with the former. Note that the highest percent reduction in variance achievable by either rerandomization or regression is  $100R^2$ , achieved with perfect covariate mean balance.

**4. Affinely invariant rerandomization criteria.** In this section we explore the theoretical implications of choosing an affinely invariant rerandomization criterion, meaning that for any affine transformation of  $\mathbf{x}$ ,  $a + \mathbf{b}\mathbf{x}$ ,  $\varphi(\mathbf{x}, \mathbf{W}) = \varphi(a + \mathbf{b}\mathbf{x}, \mathbf{W})$ . Measures based on inner products, such as Mahalanobis distance or the estimated best linear discriminant, are affinely invariant, as are criteria based on propensity scores estimated by linear logistic regression [Rubin and Thomas (1992)]. Results in this section parallel those for affinely invariant matching methods [Rubin and Thomas (1992)].

In the previous sections, we regarded  $\mathbf{x}$  as fixed, and only the randomization vector,  $\mathbf{W}$  was random. In this section, to use ellipsoidal symmetry of  $\mathbf{x}$ , we regard both  $\mathbf{x}$  and  $\mathbf{W}$  as random, so expectations are over repeated draws of  $\mathbf{x}$  and repeated randomizations.

**THEOREM 4.1.** *If  $\varphi$  is affinely invariant, and if  $\mathbf{x}$  is ellipsoidally symmetric, then*

$$(20) \quad \mathbb{E}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C \mid \varphi = 1) = \mathbb{E}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C) = \mathbf{0} \quad \text{and}$$

$$(21) \quad \text{cov}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C \mid \varphi = 1) \propto \text{cov}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C).$$

**PROOF.** First, by ellipsoidal symmetry there is an affine transformation of  $\mathbf{x}$  to a canonical form with mean (center) zero and covariance (inner product)  $\mathbf{I}$ , the  $k$ -dimensional identity matrix. The distribution of the matrix  $\mathbf{x}$  in the treated group of size  $np_w$  and the control group of size  $n(1 - p_w)$  are both independent and identically distributed samples from this zero centered spherical distribution. Any affinely invariant rule for selecting subsets of treated and control units will be a function of affinely invariant statistics in the treatment and control groups that are also zero-centered spherically symmetric. Applying  $\varphi$  creates concentric zero-centered sphere(s) that partition the space of these statistics into regions where  $\varphi = 1$  and  $\varphi = 0$ , and therefore the distribution of such statistics remains zero-centered

and spherically symmetric. Transforming back to the original form completes the proof.  $\square$

**COROLLARY 4.2.** *If  $\varphi$  is affinely invariant and if  $\mathbf{x}$  is ellipsoidally symmetric, then rerandomization leads to unbiased estimates of any linear function of  $\mathbf{x}$ .*

Note that, unlike Corollary 2.2, Corollary 4.2 applies no matter how the sample sizes are chosen.

**COROLLARY 4.3.** *If  $\varphi$  is affinely invariant and if  $\mathbf{x}$  is ellipsoidally symmetric, then*

$$(22) \quad \text{cor}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C \mid \varphi = 1) = \text{cor}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C).$$

One possible method of rerandomization, suggested by Moulton (2004), Maclure et al. (2006), Bruhn and McKenzie (2009) and Cox (2009), is to place bounds separately on each entry of  $\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C$  and ensure that each covariate difference is within its specified caliper. However, this method is not affinely invariant and will generally destroy the correlational structure of  $\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C$ , even when  $\mathbf{x}$  is ellipsoidally symmetric.

Analogous to “Equal Percent Bias Reducing” (EPBR) matching methods [Rubin (1976)], a rerandomization method is said to be “Equal Percent Variance Reducing” (EPVR) if the percent reduction in variance is the same for each covariate.

**COROLLARY 4.4.** *If  $\varphi$  is affinely invariant and if  $\mathbf{x}$  is ellipsoidally symmetric, then rerandomization is EPVR for  $\mathbf{x}$  and any linear function of  $\mathbf{x}$ .*

Rerandomization methods that are not affinely invariant could increase the variance of some linear combinations of covariates [Rubin (1976)].

Although affinely invariant methods have desirable properties in general, they are not always preferred. For example, if covariates are known to vary in importance, a rerandomization method that is not EPVR may be more desirable, allowing greater percent reduction in variance for more important covariates. Rerandomization criteria that take into account covariates of varying importance are discussed in Lock [(2011), Chapter 4].

## 5. Discussion.

5.1. *Alternatives for balancing covariates.* Rerandomization is certainly not the only way to balance covariates before the experiment.

With only a few categorical covariates, simple blocking can successfully balance all covariates, and there is no need for rerandomization. With many covariates

each taking on many values, however, blocking on all covariates can be impossible, and in this case we recommend blocking on the most important covariates, and rerandomizing to balance the components of the covariates orthogonal to the blocks. Blocking and rerandomization can, and we feel should, be used together. Multivariate matching [Greevy et al. (2004); Rubin (2006); Ho et al. (2007); Imai, King and Nall (2009); Xu and Kalbfleisch (2010)] is a special case of blocking that can better handle many covariates.

Restricted (or constrained) randomization [Yates (1948); Grundy and Healy (1950); Youden (1972); Bailey (1983)] restricts the set of acceptable randomizations in a way that preserves the validity of asymptotic-based distributional methods of analysis. However, most work on restricted randomization is specific to agricultural plots, and apparently has not been extended to multiple covariates. Blocking, matching and restricted randomization can all also be implemented through rerandomization by specifying the set of acceptable randomizations through  $\varphi$ .

The Finite Selection Model (FSM) [Morris (1979); Morris and Hill (2000)] provides balance for multiple covariates, but provides a fixed amount of balance in a fixed amount of computational time. Rerandomization has the flexibility to choose the desired tradeoff between balance and computational time. More details comparing FSM with rerandomization are in [Lock (2011), Section 5.5].

Covariate-adaptive randomization schemes [Efron (1971); White and Freedman (1978); Pocock and Simon (1975); Pocock (1979); Simon (1979); Birkett (1985); Aickin (2001); Atkinson (2002); Scott et al. (2002); McEntegart (2003); Rosenberger and Sverdlov (2008)] are designed for clinical trials with sequential treatment allocation over extended periods of time. Rerandomization as proposed here is not applicable to sequential allocation, and instead readers interested in such trials can refer to the above sources.

If covariates are not balanced before the experiment, post-hoc methods such as regression adjustment are commonly used, which rely on assumptions that often cannot be verified [Tukey (1993); Freedman (2008)]. Moreover, unlike post-hoc methods, rerandomization is conducted entirely at the design stage, and so cannot be influenced by outcome data. Tukey (1993) and Rubin (2008b) give convincing reasons for why as much as possible should be done in the design phase of an experiment, before outcome data are available, rather than in the analysis stage when the researcher has the potential to bias the results, consciously or unconsciously.

*5.2. Extensions and additional considerations.* For multiple treatment groups, any of the test statistics commonly used in multivariate analysis of variance (MANOVA) can be used to measure balance. The standard statistics are all equivalent to Mahalanobis distance in the special case of two groups. Extensions for multiple treatment groups are discussed in Lock [(2011), Section 5.2].

For unbiased estimates using rerandomization with treatment groups of unequal sizes, multiple treatment groups of equal size can be created, and then merged as needed after the rerandomization procedure, but before the physical experiment. If

extra units are discarded to form equal sized treatment groups and rerandomization is employed, precision can actually increase if covariates are highly correlated with the outcome [Lock (2011), Section 5.3].

In a Bayesian analysis, as long as all covariates relevant to  $\varphi(\mathbf{x}, \mathbf{W})$  are conditioned on, the design is ignorable [Rubin (1978)], and theoretically, the analysis can proceed as usual.

**6. Conclusion.** Randomization balances covariates across treatment groups, but only on average, and in any one experiment covariates may be unbalanced. Rerandomization provides a simple and intuitive way to improve covariate balance in randomized experiments.

To perform rerandomization, a criterion determining whether a randomization is acceptable needs to be specified. For unbiasedness, this rule needs to be symmetric regarding the treatment groups. If the criterion is affinely invariant, then for ellipsoidally symmetric distributions, balance improvement will be the same for all covariates (and all linear combinations of the covariates), and correlations between covariate differences in means will be maintained. One such criterion is to rerandomize whenever Mahalanobis distance exceeds a certain threshold.

When the covariates are correlated with the outcome, rerandomization increases precision. If the analysis reflects the rerandomization procedure, this leads to more precise estimates, more powerful tests and narrower confidence intervals.

APPENDIX

PROOF OF THEOREM 3.1. Because  $M \sim \chi_k^2$  under pure randomization when the covariate means are normally distributed, rerandomization affects the mean of  $M$  as follows:

$$\begin{aligned}
 \mathbb{E}(M \mid \mathbf{x}, M \leq a) &= \frac{(1/(\Gamma(k/2)2^{k/2})) \int_0^a y^{k/2} e^{-y/2} dy}{(1/(\Gamma(k/2)2^{k/2})) \int_0^a y^{k/2-1} e^{-y/2} dy} \\
 (23) \qquad &= \frac{\int_0^a (y/2)^{k/2} e^{-y/2} dy}{(1/2) \int_0^a (y/2)^{k/2-1} e^{-y/2} dy} \\
 &= 2 \times \frac{\gamma(k/2 + 1, a/2)}{\gamma(k/2, a/2)}.
 \end{aligned}$$

To prove (8), we convert the covariates to canonical form [Rubin and Thomas (1992)]. Let  $\Sigma = \text{cov}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C \mid \mathbf{x})$ , and define

$$(24) \qquad \mathbf{Z} \equiv \Sigma^{-1/2}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C),$$

where  $\Sigma^{-1/2}$  is the Cholesky square root of  $\Sigma^{-1}$ , so  $\Sigma^{-1/2'} \Sigma^{-1/2} = \Sigma^{-1}$ . By the assumption of normality,

$$\mathbf{Z} \mid \mathbf{x} \sim N_k(\mathbf{0}, \mathbf{I}).$$

Due to normality, uncorrelated implies independent and thus the elements of  $\mathbf{Z}$  are independent and identically distributed (i.i.d.) standard normals. Therefore, the elements of  $\mathbf{Z}$  are exchangeable.

By (5),  $M = \mathbf{Z}'\mathbf{Z} = \sum_{j=1}^k Z_j^2$ . Therefore for each  $j$  we have

$$\begin{aligned} \text{var}(Z_j \mid \mathbf{x}, M \leq a) &= \mathbb{E}(Z_j^2 \mid \mathbf{x}, M \leq a) \\ (25) \qquad \qquad \qquad &= \frac{\mathbb{E}(M \mid \mathbf{x}, M \leq a)}{k} \\ &= \frac{2}{k} \times \frac{\gamma(k/2 + 1, a/2)}{\gamma(k/2, a/2)} \\ (26) \qquad \qquad \qquad &= v_a, \end{aligned}$$

where (25) follows from the exchangeability of the elements of  $\mathbf{Z}$ .

After enforcing  $M \leq a$ , the elements of  $\mathbf{Z}$  are no longer independent (they will be negatively correlated in magnitude), but with signs they remain uncorrelated due to symmetry:

$$\begin{aligned} \text{cov}(Z_i, Z_j \mid \mathbf{x}, M \leq a) &= \mathbb{E}(Z_i Z_j \mid \mathbf{x}, M \leq a) \\ &\quad - \mathbb{E}(Z_i \mid \mathbf{x}, M \leq a)\mathbb{E}(Z_j \mid \mathbf{x}, M \leq a) \\ (27) \qquad \qquad \qquad &= \mathbb{E}(\mathbb{E}(Z_i Z_j \mid Z_j, \mathbf{x}, M \leq a) \mid \mathbf{x}, M \leq a) - 0 \\ &= \mathbb{E}(Z_j \mathbb{E}(Z_i \mid Z_j, \mathbf{x}, M \leq a) \mid \mathbf{x}, M \leq a) \\ (28) \qquad \qquad \qquad &= \mathbb{E}(Z_j \times 0 \mid \mathbf{x}, M \leq a) \\ (29) \qquad \qquad \qquad &= 0, \end{aligned}$$

where (27) follows from Corollary 2.2, and (28) follows because  $(Z_i \mid Z_j, M \leq a) \sim (-Z_i \mid Z_j, M \leq a)$ , thus  $\mathbb{E}(Z_i \mid Z_j, M \leq a) = 0$  for all  $i, j$ .

Thus after rerandomization the covariance matrix of  $\mathbf{Z}$  is  $v_a \mathbf{I}$ , hence

$$\begin{aligned} \text{cov}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C \mid \mathbf{x}, M \leq a) &= \text{cov}(\mathbf{\Sigma}^{1/2} \mathbf{Z} \mid \mathbf{x}, M \leq a) \\ &= \mathbf{\Sigma}^{1/2} \text{cov}(\mathbf{Z} \mid \mathbf{x}, M \leq a) \mathbf{\Sigma}^{1/2'} \\ &= v_a \mathbf{\Sigma} \\ &= v_a \text{cov}(\bar{\mathbf{X}}_T - \bar{\mathbf{X}}_C \mid \mathbf{x}). \qquad \square \end{aligned}$$

**Acknowledgments.** We appreciate the extraordinarily helpful comments of the editor, Professor Bühlmann, and two reviewers.

REFERENCES

AICKIN, M. (2001). Randomization, balance, and the validity and efficiency of design-adaptive allocation methods. *J. Statist. Plann. Inference* **94** 97–119. MR1820173

- ANSCOMBE, F. J. (1948a). The validity of comparative experiments. *J. Roy. Statist. Soc. Ser. A.* **111** 181–211. [MR0030181](#)
- ARNOLD, G. C. (1986). Randomization: A historic controversy. In *The Fascination of Statistics* (R. J. Brook, G. C. Arnold, T. H. Hassard and R. M. Pringle, eds.) 231–244. CRC Press, Boca Raton, FL.
- ATKINSON, A. C. (2002). The comparison of designs for sequential clinical trials with covariate information. *J. Roy. Statist. Soc. Ser. A* **165** 349–373. [MR1904822](#)
- BAILEY, R. A. (1983). Restricted randomization. *Biometrika* **70** 183–198. [MR0742988](#)
- BAILEY, R. A. (1986). Randomization, constrained. *Encyclopedia of Statistical Sciences* **7** 519–524.
- BAILEY, R. A. and ROWLEY, C. A. (1987). Valid randomization. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **410** 105–124.
- BIRKETT, N. J. (1985). Adaptive allocation in randomized controlled trials. *Control Clin Trials* **6** 146–155.
- BRILLINGER, D., JONES, L. and TUKEY, J. (1978). *The Management of Weather Resources II: The Role of Statistics in Weather Resources Management*. US Government Printing Office, Washington, DC.
- BRUHN, M. and MCKENZIE, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics* **1** 200–232.
- COCHRAN, W. G. and RUBIN, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā Ser. A* **35** 417–446.
- COX, D. R. (1982). Randomization and concomitant variables in the design of experiments. In *Statistics and Probability: Essays in Honor of C. R. Rao* 197–202. North-Holland, Amsterdam. [MR0659470](#)
- COX, D. R. (2009). Randomization in the Design of Experiments. *International Statistical Review* **77** 415–429.
- EFRON, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58** 403–417. [MR0312660](#)
- ERDŐS, P. and RÉNYI, A. (1959). On the central limit theorem for samples from a finite population. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **4** 49–61. [MR0107294](#)
- FISHER, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain* **33** 503–513.
- FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.
- FREEDMAN, D. A. (2008). On regression adjustments to experimental data. *Adv. in Appl. Math.* **40** 180–193. [MR2388610](#)
- GARTHWAITE, P. H. (1996). Confidence intervals from randomization tests. *Biometrics* 1387–1393.
- GOSSET, W. J. (1938). Comparison between balanced and random arrangements of field plots. *Biometrika* **29** 363.
- GREENBERG, B. G. (1951). Why randomize? *Biometrics* **7** 309–322.
- GREEVY, R., LU, B., SILBER, J. H. and ROSENBAUM, P. (2004). Optimal multivariate matching before randomization. *Biostatistics* **5** 263–275.
- GRUNDY, P. M. and HEALY, M. J. R. (1950). Restricted randomization and quasi-Latin squares. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **12** 286–291.
- HÁJEK, J. (1960). Limiting distributions in simple random sampling from a finite population. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5** 361–374. [MR0125612](#)
- HANSEN, B. B. and BOWERS, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statist. Sci.* **23** 219–236. [MR2516821](#)
- HARVILLE, D. A. (1975). Experimental randomization: Who needs it? *Amer. Statist.* 27–31.
- HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* **15** 199–236.
- HOLSCHUH, N. (1980). Randomization and design: I. In *R. A. Fisher: An Appreciation* (S. E. Fienberg and D. V. Hinkley, eds.). *Lecture Notes in Statistics* **1** 35–45. Springer, New York.

- IMAI, K., KING, G. and STUART, E. A. (2008). Misunderstanding between experimentalists and observationalists about causal inference. *J. Roy. Statist. Soc. Ser. A* **171** 481–502. [MR2427345](#)
- IMAI, K., KING, G. and NALL, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statist. Sci.* **24** 29–53. [MR2561126](#)
- KEELE, L., MCCONNAUGHY, C., WHITE, I., LIST, P. M. E. M. and BAILEY, D. (2009). Adjusting experimental data. In *Experiments in Political Science Conference*.
- KEMPTHORNE, O. (1955). The randomization theory of experimental inference. *J. Amer. Statist. Assoc.* **50** 946–967. [MR0071696](#)
- KEMPTHORNE, O. (1986). Randomization II. *Encyclopedia of Statistical Sciences* **7** 519–524.
- KRAUSE, M. S. and HOWARD, K. I. (2003). What random assignment does and does not do. *Journal of Clinical Psychology* **59** 751–766.
- LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. *Springer Texts in Statistics*. Springer, New York. [MR2135927](#)
- LOCK, K. F. (2011). Rerandomization to improve covariate balance in randomized experiments Ph.D. thesis, Harvard Univ., Cambridge, MA.
- MACLURE, M., NGUYEN, A., CARNEY, G., DORMUTH, C., ROELANTS, H., HO, K. and SCHNEEWEISS, S. (2006). Measuring prescribing improvements in pragmatic trials of educational tools for general practitioners. *Basic & Clinical Pharmacology & Toxicology* **98** 243–252.
- MANLY, B. F. J. (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 3rd ed. Chapman & Hall/CRC, Boca Raton, FL. [MR2257066](#)
- MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1980). *Multivariate Analysis*. Academic Press, London. [MR0560319](#)
- MCENTEGART, D. J. (2003). The pursuit of balance using stratified and dynamic randomization techniques: An overview. *Drug Information Journal* **37** 293–308.
- MORRIS, C. (1979). A finite selection model for experimental design of the health insurance study. *J. Econometrics* **11** 43–61.
- MORRIS, C. N. and HILL, J. L. (2000). The health insurance experiment: Design using the finite selection model. In *Public Policy and Statistics: Case Studies from RAND* 29–53. Springer, New York.
- MOULTON, L. H. (2004). Covariate-based constrained randomization of group-randomized trials. *Clin Trials* **1** 297–305.
- POCOCK, S. J. (1979). Allocation of patients to treatment in clinical trials. *Biometrics* **35** 183–197.
- POCOCK, S. J. and SIMON, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* **31** 103–115.
- RAYNOR, A. A. (1986). Some Sidelights on Experimental Design. In *The Fascination of Statistics* (R. J. Brook, G. C. Arnold, T. H. Hassard and R. M. Pringle, eds.) 245–264. CRC Press, Boca Raton, FL.
- ROSENBERGER, W. F. and LACHIN, J. M. (2002). *Randomization in Clinical Trials: Theory and Practice*. Wiley, New York. [MR1914364](#)
- ROSENBERGER, W. F. and SVERDLOV, O. (2008). Handling covariates in the design of clinical trials. *Statist. Sci.* **23** 404–419. [MR2483911](#)
- RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66** 688.
- RUBIN, D. B. (1976). Multivariate matching methods that are equal percent bias reducing. I. Some examples. *Biometrics* **32** 109–120. [MR0400555](#)
- RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. [MR0472152](#)
- RUBIN, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.

- RUBIN, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge Univ. Press, Cambridge. [MR2307965](#)
- RUBIN, D. B. (2008a). Comment: The design and analysis of gold standard randomized experiments. *J. Amer. Statist. Assoc.* **103** 1350–1353. [MR2655717](#)
- RUBIN, D. B. (2008b). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–804. [MR2516795](#)
- RUBIN, D. B. and THOMAS, N. (1992). Affinely invariant matching methods with ellipsoidal distributions. *Ann. Statist.* **20** 1079–1093. [MR1165607](#)
- SAVAGE, L. J. (1962). *The Foundations of Statistical Inference*. Methuen & Co. Ltd., London. [MR0146908](#)
- SCOTT, N. W., MCPHERSON, G. C., RAMSAY, C. R. and CAMPBELL, M. K. (2002). The method of minimization for allocation to clinical trials. a review. *Control Clinical Trials* **23** 662–674.
- SEIDENFELD, T. (1981). Levi on the dogma of randomization in experiments. In *Henry E. Kyburg, Jr. & Isaac Levi* (R. J. Bogdan, ed.) 263–291. Springer, Berlin.
- SIMON, R. (1979). Restricted randomization designs in clinical trials. *Biometrics* **35** 503–512.
- SOARES, J. F. and WU, C. F. J. (1985). Optimality of random allocation design for the control of accidental bias in sequential experiments. *J. Statist. Plann. Inference* **11** 81–87. [MR0783375](#)
- SPLAWA-NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. [MR1092986](#)
- SPROTT, D. A. and FAREWELL, V. T. (1993). Randomization in experimental science. *Statist. Papers* **34** 89–94. [MR1221524](#)
- TUKEY, J. W. (1993). Tightening the clinical trial. *Control Clin Trials* **14** 266–285.
- URBACH, P. (1985). Randomization and the design of experiments. *Philos. Sci.* **52** 256–273. [MR0804567](#)
- WHITE, S. J. and FREEDMAN, L. S. (1978). Allocation of patients to treatment groups in a controlled clinical study. *British Journal of Cancer* **37** 849.
- WORRALL, J. (2010). Evidence: Philosophy of science meets medicine. *J. Eval. Clin. Pract.* **16** 356–362.
- XU, Z. and KALBFLEISCH, J. D. (2010). Propensity score matching in randomized clinical trials. *Biometrics* **66** 813–823. [MR2758217](#)
- YATES, F. (1939). The comparative advantages of systematic and randomized arrangements in the design of agricultural and biological experiments. *Biometrika* **30** 440.
- YATES, F. (1948). Contribution to the discussion of “The validity of comparative experiments” by FJ Anscombe. *J. Roy. Statist. Soc. Ser. A* **111** 204–205.
- YOU DEN, W. J. (1972). Randomization and experimentation. *Technometrics* **14** 13–22.

DEPARTMENT OF STATISTICS  
DUKE UNIVERSITY  
BOX 90251  
DURHAM, NORTH CAROLINA 27708  
USA  
E-MAIL: [kari@stat.duke.edu](mailto:kari@stat.duke.edu)

STATISTICS DEPARTMENT  
HARVARD UNIVERSITY  
1 OXFORD ST.  
CAMBRIDGE, MASSACHUSETTS 02138  
USA  
E-MAIL: [rubin@stat.harvard.edu](mailto:rubin@stat.harvard.edu)